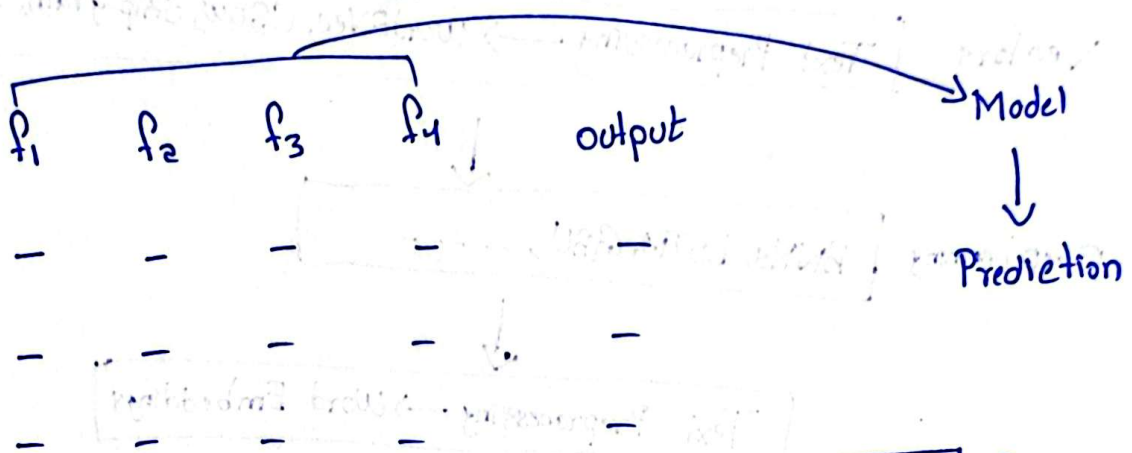
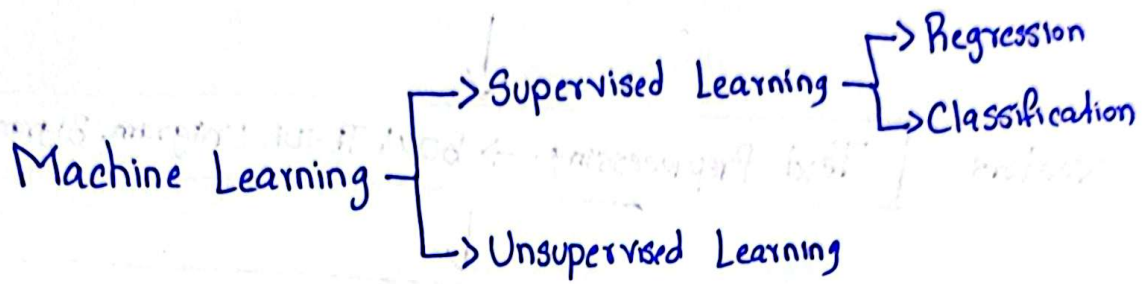
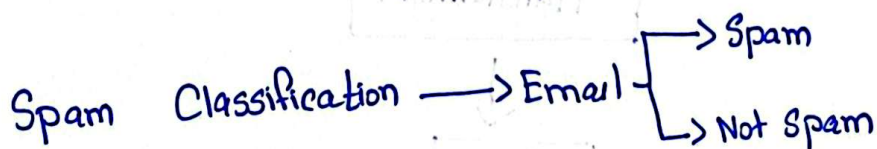


NATURAL LANGUAGE PROCESSING

→ Roadmap



Text ?



Email Subject

Email Body

Spam/Ham

↓
Vectors
(Convert)

↓
Convert to
vectors

Cleaning

Text Preprocessing → Tokenization, Lemmatization, Stemming

Vectors

Text Preprocessing → BOW, TF-IDF, Unigram, Bigram, ...

Vectors

Text Preprocessing → Word2Vec, CBow, Skip-gram, ...

Deep Learning

RNNs, LSTM, GRU, ...

Text Preprocessing → Word Embeddings

Transformers

BERT

→ Use Cases

① Spelling correction

⑥ Question Answering

② Translation

⑦ Google Assistant

③ Text to images/videos

④ Summarization

⑤ Classification (Text)

→ Tokenization

Tokenization is a process of converting a paragraph into sentences and sentences into words

- ① Corpus → Paragraph
- ② Documents → Sentences
- ③ Vocabulary → Unique words
- ④ words → words

Corpus { "My name is Saad and I have a interest in Machine Learning." "I am also a Data Science Consultant." }



Tokens {sentences}

- ① My name is Saad and I have a interest in Machine Learning
- ② I am also a Data Science Consultant



Tokenization {words}

- ① My, name, is, Saad, and, I, have, a, interest, in, Machine, Learning
- ② I, am, also, a, Data, Science, Consultant

→ Stemming and Lemmatization

> Stemming

It is the process of removing the last few characters of a given word, to obtain a shorter form, even if that form doesn't have any meaning

History
Historical

History

Finally
Final

Final

> Lemmatization

It is a linguistic process that involves reducing words to their base or root form known as lemma.

History
Historical

History

Finally
Final

Final